Building personalised synthetic voices for individuals with severe speech impairment

Sarah Creer, Stuart Cunningham, Phil Green, and Junichi Yamagishi. Building personalised synthetic voices for individuals with severe speech impairment. Computer Speech & Language, 27(6):1178 - 1193, 2013. Special Issue on Speech and Language Processing for Assistive Technology.

Building personalised synthetic voices for individuals with severe speech impairment

Sarah Creer[a,*], Stuart Cunningham[b], Phil Green[c], Junichi Yamagishi[d]

[a]*School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK*
[b]*Department of Human Communication Sciences, University of Sheffield, 31 Claremont Crescent, Sheffield S10 2TA, UK*
[c]*Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK*
[d]*Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, UK*

Abstract

For individuals with severe speech impairment accurate spoken communication can be difficult and require considerable effort. Some may choose to use a voice output communication aid (or VOCA) to support their spoken communication needs. A VOCA typically takes input from the user through a keyboard or switch-based interface and produces spoken output using either synthesised or recorded speech. The type and number of synthetic voices that can be accessed with a VOCA is often limited and this has been implicated as a factor for rejection of the devices. Therefore, there is a need to be able to provide voices that are more appropriate and acceptable for users.

This paper reports on a study that utilises recent advances in speech synthesis to produce personalised synthetic voices for 3 speakers with mild to severe dysarthria, one of the most common speech disorders. Using a statistical parametric approach to synthesis, an average voice trained on data from several unimpaired speakers was adapted using recordings of the impaired speech of 3 dysarthric speakers. By careful selection of the speech data and the model parameters, several exemplar voices were produced for

*Corresponding author
Email addresses: S.Creer@Sheffield.ac.uk (Sarah Creer), S.Cunningham@Sheffield.ac.uk (Stuart Cunningham), P.Green@dcs.shef.ac.uk (Phil Green), jyamagis@inf.ed.ac.uk (Junichi Yamagishi)

each speaker. A qualitative evaluation was conducted with the speakers and listeners who were familiar with the speaker. The evaluation showed that for one of the 3 speakers a voice could be created which conveyed many of his personal characteristics, such as regional identity, sex and age.

Keywords:
speech synthesis, augmentative and alternative communication, disordered speech, voice output communication aid

---

1. Introduction

The use of Voice Output Communication Aids (VOCAs) has been shown to increase quality of life for individuals with speech impairment (Mathy et al., 2000). VOCAs take input from a user through a keyboard or switch-based interface and use a pre-recorded or synthesised voice as output to approximate oral communication that occurs between conversational partners. Whether people persevere with using a VOCA depends on the acceptability of the device including motivation and attitudes of both conversation partners and features of the technology itself (Mathy et al., 2000; Lasker and Bedrosian, 2001). Lasker and Bedrosian's (2001) model of acceptability of augmentative and alternative communication (AAC) interventions explicitly includes customisation of the device and voice output quality. This suggests that if the range of voices available in VOCAs does not provide a suitable choice for the user, the associated risks of abandonment are increased.

People who use VOCAs are individuals who have lost or are losing the ability to produce their own speech due to either acquired conditions such as motor neurone disease (MND), or congenital conditions such as cerebral palsy (CP).

The speech impairment due to such conditions is collectively termed dysarthria. In general, dysarthria is characterised by abnormalities in the speed, range and accuracy of movement required to control the respiratory and articulatory systems used in speech production (Duffy, 2005).

Acquired disorders are either progressive such as Parkinson's disease, or can have a sudden onset as a result of traumatic brain injury such as a stroke or cerebrovascular accident (CVA). Congenital conditions are usually stable in presentation whereas acquired progressive disorders are usually preceded by having normal speech development and the diminishing neurological function leads to a progressive deterioration in the individual's ability to produce

speech. Dysarthria is thought to affect 170 people per 100,000 in the UK (Enderby and Emerson, 1995).

In progressive conditions, deterioration of speech is usually the first symptom to present (Duffy, 2005; Holmberg et al., 1996) and as motor control is lost, the severity of impairment increases and understanding the speech becomes more difficult.

The severity of dysarthria depends on the location and extent of the brain injury, which defines the type of dysarthria, formalised by Darley et al. (1969) as: spastic, flaccid, ataxic, hypokinetic, hyperkinetic and mixed. In general terms the more severe the dysarthria the greater the reduction in intelligibility, voice quality and prosody, see Duffy (2005); Enderby (1983); Weismer (2007); Ziegler (2008).

When an individual experiences speech impairment, maintenance of social interaction is vital for the avoidance of withdrawal from society (Light, 1988; Murphy, 2004; O'Keefe et al., 1998). Using a VOCA, an individual has to like and identify with the voice to feel motivated to use it. A voice provides clues about the gender, age, size, ethnicity and geographical identity of that individual (Chambers, 1995; Wells, 1982) and is a personal identifier of an individual to family members and acquaintances. Embarrassment and negative attitudes towards an individual's own speech can create barriers to socialisation (Miller et al., 2006), which can be extended to having negative attitudes towards the voice in a VOCA that they are using.

Without personalisation of identity, an individual's ability to form associations with others through their speech may be lost, which is detrimental to their participation in society (Angelo et al., 1996; Hetzroni and Harris, 1996; Parette and Huer, 2002; Smith, 2005).

Evaluations of fluent speakers' preferences for VOCA ~~use~~ matched the most natural-sounding and gender-appropriate voice to themselves (Crabtree et al., 1990). This supports results from studies of assistive technology design, suggesting that individuals prefer a VOCA to have a voice that is consistent with the characteristics of the person who is using it (Light et al., 2007). Offering the user a choice of voices for their communication aid including one which matches their vocal identity pre-deterioration could lead to more acceptance of this type of technology.

For VOCA users, there is limited choice of voices available to distinguish themselves from others and to represent themselves. Personalisation has previously been attempted by interpolating between existing voices (Murray and Arnott, 1993) or morphing voices to more closely match the user in terms

of age, such as in the Tango (BlinkTwice) communication aid.

Another approach at personalisation can be found in the ModelTalker project (Bunnell et al., 2010). This supports a procedure to "bank" speech recordings and create a personalised synthetic voice from those recordings. This approach is specifically designed for people with progressive conditions who may make the recordings before their speech has deteriorated. To build an acceptable voice requires a large amount of data to be recorded, which can be difficult for individuals with conditions causing speech impairment to produce.

Commercial alternatives requiring large datasets are also available, for example, film critic Roger Ebert and American footballer Steve Gleason who lost their voices due to medical conditions and had personalised voices built for them by CereProc Ltd. These cases highlight the demand for personalised vocal output in a communication aid.

Those individuals with sudden onset acquired conditions such as CVA, or congenital conditions have no opportunity to personalise a synthetic voice to their own characteristics. Work carried out to address this problem has attempted to capture the relatively unimpaired source characteristics from a dysarthric speaker and use it to replace those of a fluent synthesised voice working within the ModelTalker framework (Jreige et al., 2009). However a technique that addresses both source and filter components could convey much more of the individual's vocal identity.

This paper introduces the potential benefits of providing personalised synthetic voice output and identifies a target population who are not currently being provided with access to personalised voice building, namely those whose speech has begun to deteriorate due to a progressive condition. The paper extends the work detailed in Creer et al. (2010) and provides detailed case studies for the work referred to in Yamagishi et al. (2012).

Section 2 reviews statistical parametric synthesis and section 3 proposes a methodology for applying this procedure to dysarthric speech data. Section 4 introduces the procedures involved in building voices for 3 individuals with different types and severity of dysarthria and presents the methodology for the evaluation of these voices. The results of the evaluation are presented in section 5 and discussed in section 6.

2. Statistical parametric synthesis

To provide personalised synthetic voices for VOCA a technique is required which produces intelligible and natural-sounding speech. Moreover, the speech output should be sufficiently similar to the user's own voice pre-deterioration, or in the case of users with congenital conditions be sufficiently similar to the individual characteristics of their voice. The technique must be able to be take into account data which has begun to deteriorate and not recreate the errors in the output while still using a minimal amount of data for input.

One such system is model-based statistical parametric speech synthesis which uses Hidden Markov model (HMM) based techniques to probabilistically model and generate sequences of feature vectors, discrete representations of the speech signal at a segment of time (Zen et al., 2009). Models are trained on a corpus of speech data to produce statistical representations of the acoustics. Novel speech utterances are then formed by concatenating the appropriate models, generating the most likely sequence of feature vectors from the concatenated model from which a speech waveform is synthesised.

This data-driven technique produces highly intelligible, consistent output and is more robust to inconsistent recording conditions than systems built using pre-recorded sections of speech (Yamagishi et al., 2008a). Using this system means speaker adaptation techniques can be used to adapt from robust speaker-independent models to personalise the system using a minimal amount of data.

The HTS toolkit ('H Triple S' - HMM-based Speech Synthesis System) (Yamagishi et al., 2007; Zen et al., 2007; Zen and Toda, 2005), an extension to the HTK speech recognition toolkit (Young et al., 2002), provides a research tool for HMM-based synthesis and is described in more detail in the following sections.

2.1. Feature vectors

To build the models, the speech data has to be discretised into perceptually relevant feature vectors, which are sufficiently detailed to reconstruct the speech signal sufficiently accurately to produce a natural-sounding output.

HTS uses STRAIGHT (Kawahara et al., 1999) vocoding to both extract features and resynthesise the waveform. The feature vectors comprise separate streams for: spectral features including energy; log F0, the acoustic

correlate of pitch; band aperiodicity, representing the relative energy of aperiodic components in the periodic signal and the delta and delta-delta dynamic components (Furui, 1981) of each stream.

F0 is modelled on a logarithmic scale as motivated by the Fuijisaki model (Fujisaki and Hirose, 2000) which facilitates the combination of F0 contours at both the accent and phrase levels of speech.

HTS simultaneously models the features to ensure that the alignment between the spectral features and the prosodic features remains consistent.

2.2. Models

An HMM consists of a statistical model of the observed data in the form of feature vectors and the temporal sequence in which they occur. The temporal variation of speech is modelled with a Markov chain of states with associated transition probabilities between these states. Associated with each state is a statistical model of the acoustics of a particular segment of speech, usually a continuous probability distribution. To estimate this statistical representation, a training process is performed. The model is exposed to multiple examples of the unit being modelled and its parameters are re-estimated such that the likelihood of the model, given the examples, is maximised.

The different states capture sub-phonetic temporal variation. The unit modelled in HTS is the context-dependent phone, phone-sized units with contextual information and is modelled by five emitting states. This relatively high number of states allows acoustic information to be captured with high temporal resolution.

A geometric duration distribution is implied by standard HMMs, which is a poor model of actual phone durations. HTS therefore estimates a normally distributed state duration probability density for each state in each model during training, which is explicitly attached to the model for both training and synthesis. This alters some of the mathematical properties of the model and results in a Hidden Semi-Markov Model (HSMM) (Zen et al., 2004). The training corpus or adaptation data is used to estimate the parameters of the duration model.

Rich contextual information is required to contribute to the generation of phonetic and prosodic elements of the output synthesised speech. In HTS contextual phonetic and prosodic information is provided at the phoneme, syllable, word, phrase and utterance levels. The data sparsity problem introduced by the rich contextual information is addressed by sharing the parameters of the state output distribution between acoustically similar states.

This sharing is performed using decision trees which define clusters of acoustically similar states using splitting questions based on the detailed phonetic and prosodic contextual labels. The minimum description length criterion (Shinoda and Watanabe, 2000) is used to determine both the structure and complexity of the decision tree.

Different contextual factors affect the acoustic distance between vectors for duration, spectral information, log F0 and aperiodicity and so these feature streams are clustered independently of each other. There are separate models for each feature stream and they are combined only at synthesis time.

2.3. Adaptation

For this application, an adaptation method using minimal input data is required. It is possible to adapt speaker-independent or average voice models, trained on large amounts of data from multiple speakers, to more closely match a speaker's individual voice characteristics. The average voice starting point provides a strong prior for the adaptation data, data taken from one speaker used to adapt the models, and enables robust estimation of the target speaker model. In building an average voice model, speaker- and gender-dependent characteristics in the data are neutralised capturing a robust model of the phonetic variation in speech, not a model of inter-speaker variation. This is done using speaker adaptive training (SAT) for parameter re-estimation. The SAT framework (Anastasakos et al., 1996) ensures that the acoustic variation due to the speaker population is reduced when estimating the variance of the acoustic model parameters. The adaptation procedure then transforms the average voice model towards the target speaker.

2.4. Synthesis

The first stage of synthesis is to convert the orthographic text to be synthesised into a sequence of context-sensitive labels. A composite HSMM is then created by concatenating the context-dependent models corresponding to this label sequence. A duration is assigned to each state in the composite HSMM which maximises the likelihood of the state duration probability density.

The feature sequence of maximum conditional probability, given the input state sequence and models, including the probability distributions over the deltas and delta-deltas as well as those for the static features, is found using the feature generation algorithm (Tokuda et al., 2000). This feature sequence

is subsequently converted into a waveform using the STRAIGHT vocoder (Kawahara et al., 1999).

2.5. Global variance

The statistical nature of this technique results in spectral details being averaged out with high priority placed on producing a smooth output trace for each feature. In an attempt to improve the speech output and prevent over-smoothing, refinements to the feature generation algorithm were introduced which model the utterance level variance (also called the global variance) of each stream. For each utterance in the adaptation data and for each set of features: mel cepstra, log F0 and aperiodicity, a variance is calculated. The mean of these variances and the variance of these variances is calculated across all the utterances in the data set. The global variance is integrated into the feature generation algorithm and ensures that the features generated more accurately reflect the utterance level variance of the data rather than over-smoothing the cepstral coefficients, log F0 and aperiodicity output (Toda and Tokuda, 2007).

3. Statistical parametric synthesis using impaired speech

For those speakers whose speech has begun to deteriorate, the synthesis needs to avoid reconstructing errors in production or misalignments between models and acoustics in the output. The following sections further detail the issues involved for using HTS with dysarthric data and attempt to provide a solution for reconstructing the voices of individuals, compensating for any impairment captured in adapting the models.

3.0.1. Data selection

In the adaptation stage of the HTS procedure, at each iteration, an alignment between the data and current models is performed. If the sum of all the utterance alignment likelihoods is too low (log likelihood less than -10^{10}), the whole utterance is rejected from the adaptation data along with potentially intelligible sections. Speech production is a difficult task for the target individuals and therefore a way of maximising the use of this data is required.

Intelligible sections of speech can be extracted from the recordings and associated with the corresponding sections of the full phonetic and prosodic context transcription derived from orthographic transcription of the expected

data. Although not necessarily an accurate representation of what was actually produced by the speaker and the surrounding context, this method links the speech produced with the cognitive planning of what was intended to be said, as shown through the presence of anticipatory coarticulation in the data (Katz, 2000) and avoids the need for expensive and difficult phone-level relabelling. This approach allows a much higher percentage of data to be accepted as adaptation data than if the data has not been edited, specifically for those speakers with more severe dysarthria. By using only intelligible segments of speech data, the possibility of recreating dysfluencies in the output voice is minimised.

Data selection can be done manually by a human listener making a judgement on whether a section is intelligible. Using human judgement and manual selection of intelligible data is not an ideal solution as it is time-consuming and inconsistency in judging intelligibility arises as a human becomes more attuned to the speech of an individual over time (Carmichael and Green, 2003).

Figure 1 demonstrates the complexity of trying to automate the selection process. The figure shows a short section of an Arctic database (Kominek and Black, 2003) sentence as spoken by a dysarthric speaker: "I may manage to freight a cargo back as well". The transcription panel nearest the spectrogram shows what different fragments are present in the speech files. They consist of pauses (labelled 'pau'), words or syllables (shown in forward slash delimiters '//'), non-vocal sounds (labelled 'noise') and vocal insertions (labelled 'vocal'). The central panel shows where the words occur in the phrase and the topmost panel shows which sections of the phrase are selected as being intelligible and therefore usable as adaptation data.

Detecting non-speech noise including silences and extraneous noise may be possible to automate but there is a high occurrence of vocalised noise. The vocalised insertions in the dysarthric output have speech-like characteristics which makes it more difficult to automatically discriminate them from the speech that is to be retained in the adaptation data.

3.0.2. Feature selection

If there are errors in a dysarthric individual's speech, it would be useful to only use for adaptation those features which are not affected by the disorder. The remaining affected features would not be used as target speech for adaptation but the corresponding features in the starting point average voice model would be retained. The structure of HTS allows an approxi-

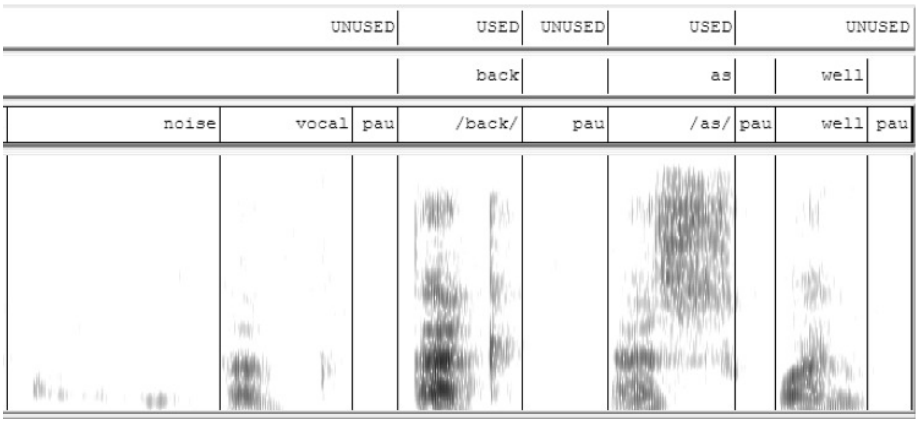

Figure 1: *The phrase "back as well" as spoken by a dysarthric speaker. It is labelled to show which sections of the phrase are usable as adaptation data (USED/UNUSED). The central pane indicates the word boundaries. The third pane shows the type of segment: pauses (pau), words or syllables (shown between //), non-vocal sounds (noise) and vocal insertions (vocal).*

mation to this behaviour. The feature vectors are extracted and used to adapt the HSMMs simultaneously, but the spectral, log F0, aperiodicity and duration features are represented in separate streams and re-combined only when generating the synthesised speech. Therefore post-adaptation, certain features of the speech can be substituted with those of the original average voice and used to reconstruct those features showing impairment. Figure 2 shows the type of substitutions that can be made using information from the average voice and the dysarthric speaker participant model to create an output speaker model. Features that capture the speaker characteristics are taken from the speaker participant model and information from the average speaker model reconstructs those features affected by the speaker's condition. Different feature substitutions depend on the individual's condition and stage of deterioration.

Spectral information: Most of the inaccurate articulation that occurs in the adaptation data is removed during the data selection process. Certain aspects of dysarthric speech such as nasalisation and distortions caused by secondary articulations may remain in the data and the extent to which this remains depends on the human judgement of how much that distortion is perceived.

The adaptation data that remains is therefore a reasonable characterisa-

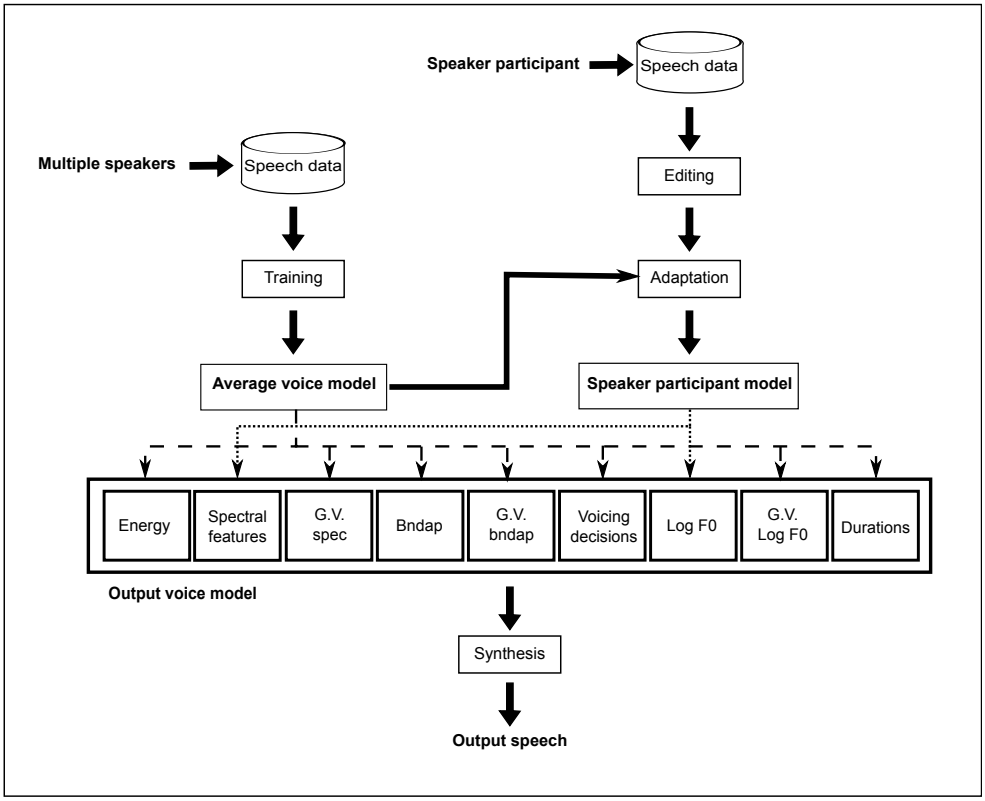

Figure 2: *Possible component feature selection to produce an output speaker model with speaker characteristics taken from the speaker model and components taken from the average voice model which compensate for the effects of the individual's condition. Bndap is band aperiodicity and G.V. is global variance. The dashed arrows show where the output voice feature models are taken from: the average voice model or the speaker participant model.*

tion of the participant's speech where the dysarthria is minimally perceived. Retaining the participant's spectral information in this way allows the retention of the individual's speaker characteristics.

Global variance for spectral features: Articulations of sounds that are intelligible may be more highly variable in dysarthric data than in typical speech. This spectral variability is modelled by the global variance parameter (see section 2.5), which influences the utterance level spectral variance during the parameter generation process. Where this variance is high, as could be the case for dysarthric data, constraining this measure could be beneficial to the output. The average voice model global variance for the spectral features can therefore be used with the speaker participant spectral features to constrain the variability and produce a more well-defined spectral output.

Energy: Energy may be highly variable due to the speech disorder. The zeroth mel cepstral coefficient can be selected from the average voice model and used in combination with the mel cepstral coefficients from the speaker participant model in the output speaker model. This smooths the output if there is much variation in the energy in the original speech and produces a more appropriate speaker energy if the speaker's voice has either reduced or elevated energy levels.

Log F0: The F0 of the speaker should be used in the output speaker models, if it has not been adversely affected by the condition, as it contains information specific to the speaker and contributes to the recognition of the voice as belonging to that particular individual.

Voicing decisions: Where there are phonatory irregularity problems such as abnormal production of voicing, voicing initiation and reduced control of the vocal folds, voicing decisions can be isolated from the average voice log F0 model and used in the output speaker log F0 model.

Global variance for log F0: Where the speaker has either a monopitch or highly variable prosodic quality due to the condition, the global variance of the log F0 can be altered to make the pitch range more appropriate. This can be done either by changing the mean of the global variance to that of the average voice or altering it to an amount which is appropriate for that speaker. This parameter can be customised to suit the preference of the speaker and how this alteration affects the intelligibility and naturalness of the synthesised speech.

Aperiodicity: The pitch periods of normal speakers are generally highly regular. However, the individual with dysarthria may have altered voice quality caused by reduced control of the larynx and weakened or tightened

vocal folds. This causes an abnormal setting of the vocal folds, either causing excessive breath through the glottis or having to force the air through the constricted glottal area, in either case producing unwanted turbulent noise in the signal. Substitution of the aperiodicity models from the average voice alters the voice quality effect to match that of the average voice.

Global variance for aperiodicity: Using the average voice global variance of the aperiodicity may also help to constrain the potentially increased variability of the aperiodicity in the speaker models caused by the individual's condition.

Duration: For dysarthric speakers, the duration of segments is highly variable and often disordered, causing distortions in the rhythm and intonation of the output. This problem is partly dealt with in the data selection process for adaptation but this selection process will not remove the variability that occurs when the speech is of varying speeds but well-articulated. By using the average voice model duration probability distributions, a consistent and reliable estimate of the duration of the segments will be produced.

Speech rate: To make the output synthesis more appropriate and preferable for the user, the speech rate can be altered during synthesis, using the average voice model relative durations as a starting point.

Table 1 summarises which aspects found in dysarthric speech can be solved by data selection and substitution of average voice model information into the speaker participant model to produce an acceptable output speaker model.

4. Method

To evaluate whether these substitutions make using HMM-based synthesis viable for building voices for individuals with dysarthric speech, an experiment evaluated voices built using these modifications. This was implemented for three individuals with different pathologies.

The evaluation posed three questions:

1. *Can the individual recognise themselves in the output voices and which features contribute to this recognition?*
2. *Which features affect the quality of the voice output for the different speaker participants?*

Problem	**Solution**
Maximising use of data available for adaptation	Data selection
Articulation problems	Data selection
Highly variable articulation accuracy	Data selection and use average voice global variance for spectral features
Highly variable intensity or intensity decay	Use average voice energy
Laryngeal voice onset problems	Use average voice voicing decisions
Incorrect voicing in segments	Use average voice voicing decisions
Reduced F0 range	Use average voice or altered global variance for log F0
Altered voice quality	Use average voice aperiodicity
Highly variable or inappropriate segment duration	Use average voice durations
Highly variable or inappropriate speech rate	Use average voice durations and alter output rate

Table 1: *Proposed solutions for reconstructing voices showing dysarthric features.*

3. *Can features be altered to make the voices more appropriate for that speaker?*

Question 1 aimed to provide information about which features should be used in the output model to capture the individual's speaker characteristics. It also aimed to provide a measure of how well the output model captures the speaker's identity in the synthesis. Question 2 aimed to see which of the possible substitution of features would affect or improve the output voice synthesis, where quality is associated with its potential practical use in a communication aid. Question 3 aimed to provide more information on the flexibility of the system in terms of alteration of the prosodic output.

The target voice in this experiment is not a tangible recording of the speech pre-deterioration. It is defined as a voice which is recognisable as the original speaker but reconstructed to provide an intelligible synthesised voice without the dysarthric features. To evaluate whether the output speech is appropriate for that speaker for potential use in a communication aid, the participants should be able to make a judgement without hearing target speech for comparison. For this reason, the evaluation takes a qualitative

approach using as participants the speakers themselves (speaker participants) and people who know the speakers (listener participants).

4.1. Participants

4.1.1. Speaker participants

The speaker participants reported in this study are identified as speakers 1, 2 and 3 and are all British English speakers. Speaker 1 was male, aged 80 years old at the time of recording, two years post cerebrovascular accident (CVA), with moderate flaccid dysarthria. In his speech overall energy varied, with imprecise and slow movement of the articulators resulting in a slow rate of production. Speaker 2 was male, aged 69 years old at the time of recording and had been diagnosed with Parkinson's disease six years previously. He showed symptoms of mild hypokinetic dysarthria. His speech was quiet, with variable energy. There was little variation in pitch and a high perceived rate of articulation. Speaker 3 was male, aged 80 years old at the time of recording with severe primary progressive apraxia of speech and dysphasia, which had onset six years previously. His dysarthria was classed as moderate at the time of recording. His speech was telegraphic, it contained many insertions, with imprecise and slow movement of the articulators resulting in a slow rate of production.

These speakers show a range of different pathologies and severity of dysarthria. The speakers were all male due to the lack of availability at the time of a female average voice model. The degradation of quality of the synthesised speech output when adapting from incorrectly matching gender-dependent models rather than gender-appropriate models (Isogai et al., 2005) would have added further complexity to the evaluation. To fully investigate the possibilities of using HTS for voice building for speakers with dysarthria a much wider population and many more speakers would be required to make quantitative and statistically significant claims.

The evaluation was conducted using speaker participants 1 and 2 evaluating their own voice. Speaker 3 did not participate in the evaluation due to his condition at the time.

4.1.2. Listener participants

Two listeners who knew speaker 1 and two who knew speaker 2 evaluated their voices and one listener participant evaluated speaker 3's synthetic voices. The listener participants were student or staff members of the University of Sheffield. They were speech and language therapists not only par-

ticipating as listeners familiar with the speakers but also as expert listeners. The listeners reported no hearing or speech impairment and were not paid for their participation.

4.2. Data collection

Data was collected for speakers 1 and 2 in a quiet clinic room in the Department of Human Communication Sciences, University of Sheffield using a Morantz PMD670 audio recorder with a Shure SM80 microphone. For practical reasons due his condition, speaker 3's data was recorded onto a laptop computer using the internal microphone in his own home.

The recorded material was taken from the Arctic dataset A. The sentences were presented to speakers 1 and 2 on separate sheets of paper in a folder to avoid any listing intonation effects in the reading and to maintain consistent recording conditions. The participants were asked to read the sentences as naturally as possible. The participants completed the recordings in one sitting but were encouraged to take breaks with a drink of water at least every 50 sentences or as often as they felt necessary. In these conditions, speaker 1 recorded the first 200 sentences of the Arctic set A and speaker 2 recorded the first 150 of the same set.

For speaker 3, the sentence prompts were displayed one at a time on the computer screen using Prorec 1.01 Speech Prompt and Record system (Huckvale, 2009). The sentences were recorded in sections of 20 per session and the aim was to record two recording sessions a day, however, it was left to the speaker to decide when during the day he felt able and motivated to do the recordings. Speaker 3 completed the recording of the first 379 of the set A sentences.

4.3. Building voices

The voices were built using HTS version 2.1 (internal) with 138-dimensional feature vectors, capturing the static and dynamic components of the signal. The spectral stream is 120-dimensional, consisting of 40 STRAIGHT mel cepstra (including energy), their deltas and delta-deltas. The log F0 is represented in 3 dimensions: log F0, deltas and delta-deltas. The band aperiodicity component is 15-dimensional consisting of 5 different frequency band representations: 0-1, 1-2, 2-4, 4-6 and 6-8 kHz, deltas and delta-deltas. The feature vectors were extracted from the speech every 5 ms with ~~a window size of 25 ms.~~

The average voice was built from full Arctic data sets (1132 sentences) as spoken by 6 male speakers: 4 US English speakers, 1 Canadian English speaker and 1 Scottish English speaker. At the time of the study, this was the only pre-built average voice available to use with the toolkit and so was used for expediency.

The adaptation was done using a combination of constrained structural maximum a posteriori linear regression (CSMAPLR) (Yamagishi et al., 2009) and maximum a posteriori (MAP) techniques.

For each speaker, two voices were built: one with all the unedited data the participants had recorded and the other with data manually selected for intelligibility (see figure 2 for more details on the process). Any sections with noise, unlabelled pauses or articulations which were unintelligible or did not match the labels derived from the original prompts were removed from the adaptation data. The edited voices used were built with segments totalling the equivalent of 172 sentences for speaker 1, 119 sentences for speaker 2 and 182 sentences for speaker 3. Equivalence is based on the average sentence length of Arctic set A being 8.9 words (correct to 1 decimal place) (Kominek and Black, 2003).

4.4. Stimuli

The stimuli presented to the participants were synthesised sentences and paragraphs taken from SCRIBE (Spoken Corpus Recordings in British English) (Huckvale, 2004). The SCRIBE paragraphs contain a high frequency of words which have features attributable to different regional accents of British English. The passages were chosen to be long enough for the listener to get a general impression of the features of the voice without focussing on individual errors and keeping the evaluation to an appropriate length.

4.5. Procedure: speaker participants

The evaluations took place in a quiet room at the University of Sheffield. The stimuli were presented to the participants individually using a laptop computer with external speakers.

The research was introduced as building voices for a computer to use to speak for that individual on days where their own voice was not clear. An example of the average voice was introduced as a starting point from which the voice was changed to an approximation of the participant's voice, based on the data that they recorded previously. An original recording of two non-disordered voices built with 500 sentences was played, followed by the

synthesised version of the same sentence and a sentence for which the listeners had not heard an original recording. This was to make the participants aware of the potential of this system. For each voice the participants were asked to rate the similarity of the synthesised output to the original recordings on a 1 (sounded like a different person) - 5 (sounded like the same person) scale. This attempted to gauge their reaction to the synthesised voices whilst starting to attune their hearing to synthesised speech.

An example of their own speech from the original recording was played to the participant to make them aware of the sound of their own voice as played through the speakers, attempting to address the problem of hearing one's own speech through a different medium of sound transmission (von Bekesy, 1949).

4.5.1. Similarity to target

Comparisons were made between the average voice and voices synthesised with average voice components introducing features that display speaker characteristics (van Dommelen, 1990) taken from the speaker participant model. The voices were built using edited data only. Conditions in this stage of the evaluation were: average voice, average voice with participant log F0 features, average voice with participant spectral information and average voice with participant log F0 and spectral information, see figure 2 for how the voice features were combined. Participants were asked to rate the difference between the original recording and the synthesis on a 1 (does not sound like me) - 5 (sounds like me) scale, with only the end-points labelled.

The same paragraph was played for each condition, the content of which was not part of the set of original recordings used to build the voices.

4.5.2. Output quality

The second evaluation question addressed the overall quality of output, combining intelligibility and naturalness with similarity of speaker. A choice was presented between the average voice with participant spectral and log F0 features (which is the configuration used in figure 2) and the same voice with one additional feature of the speaker participant's model substituted. The question asked was "For each pair, which voice do you think sounds best?", allowing the individual to have their own criteria for their definition of "best". Conditions evaluated were: use of the participant's energy, use of the participant's durations, use of the participant's global variance for spectral features (all using the set of edited data) and the final condition

used the full set of unedited data with only participant spectral and log F0 features to build the voice. These conditions were chosen for evaluation as they had a perceived effect on the output for at least one of the participants. The participant could indicate that they perceived no difference between the two samples. The pairs were randomly ordered and could be listened to as many times as was required. One paragraph was used for each condition.

4.5.3. Appropriateness for speaker

The third evaluation question dealt with appropriateness of synthetic speech output for that participant and their preferences for the customisable features: speech rate of utterance and global variance for log F0. A pairwise comparison was made for three different sentences. For rate, the comparison was between the average voice durations and a slowed down version of the average voice durations. For global variance for log F0, the two options were that of the average voice or that calculated from the participant's adaptation data. For each pair the question was asked "Can you tell a difference and if so, which one do you prefer?".

Follow up questions to access the overall acceptability of the voice were then posed as follows:

- Do you like the voice? For the one you liked the best, can you give a rating of 1 (do not like the voice) - 5 (like the voice)?
- On days when you felt your voice was not clear, would you be happy to use that synthesised voice instead?
- If you could choose between using this voice or an alternative voice (an example of a commercially available voice from Acapela (Peter)), which would you prefer?

4.6. Procedure: listener participants

The procedure for the listener participants experiment closely followed that designed for the speaker participants. For the listeners who knew speakers 1 (listeners 1A and 1B) and 2 (listeners 2A and 2B), the stimuli were presented to both participants at the same time in each evaluation to allow for discussion although their responses were recorded separately.

The listeners were presented with only one of the non-disordered speech voices. They were not presented with original recordings from the speaker's data set allowing responses to the stimuli based only on their perception of

whether the output could be associated with the speaker themselves rather than a direct measure of similarity to the initial recordings.

The questions asked during the presentation were for section 1 "Does this voice sound like the speaker?", for section 2 "Which of these voices sounds best for the speaker?" and for section 3 "Can you tell a difference and if so, which one is most appropriate for that speaker?". Follow up questions were not asked to the listener participants.

5. Results

5.1. Preliminary question

The results of the preliminary part of the experiment for speakers and listeners are displayed in table 2. In this part of the experiment both the speakers and listeners were asked to rate how similar the synthetic voice sounded to recordings of the original speaker. The voices were constructed from recordings of typically-speaking individuals and indicate the listener's ability to assess the similarity of the original and synthetic speech.

	Voice 1
Speaker 1	5
Listener 1A	4.5
Listener 1B	4.5
Speaker 2	1
Listener 2A	4
Listener 2B	4
Listener 3A	4

Table 2: *Ratings from speakers and listeners evaluating a voice built from non-disordered data on a 1(does not sound like that speaker) - 5(sounds like that speaker) similarity scale.*

The results show that the listeners and one of the speakers, rate the synthetic voice as being similar to the original speech.

5.2. Similarity to target

The results for the first question in the formal evaluation are shown in table 3. After exposure to the stimuli, speaker 1's rating of the average voice was high, showing that he perceived the average voice as sounding similar to his own. The rating increased to 5 for all other conditions containing components of his model substituted into the output voice. Listeners 1A and 1B note more discriminating differences between the voices, agreeing that introducing speaker log F0 alone is insufficient to recognise speaker characteristics in the output synthesis. The similarity increases as the speaker's own spectral features are used and further increases when using both speaker spectral features and log F0.

Speaker 2's ratings showed that he did not recognise himself in the voice. The similarity rating was not high from the listeners but there was more recognition of the speaker when information taken from the speaker's models was introduced.

For speaker 3, the listener judged that there was no similarity of the voices to the speaker until both the speaker spectral features and log F0 were introduced when the rating slightly increased.

Voice	Ave	Ave+sp logF0	Ave+sp mel cep	Ave+sp logF0+mcep
Speaker 1	4	5	5	5
Listener 1A	1	1	2	3
Listener 1B	1	1	1.5	3
Speaker 2	1	1	1	1
Listener 2A	1	1	2	2
Listener 2B	1	1	2	2
Listener 3A	1	1	1	2

Table 3: *Ratings from speakers and listeners evaluating voices built from average voice models with different speaker participant model components introduced. Ratings are on a 1(does not sound like me/him) - 5(sounds like me/him) similarity scale.*

5.3. Output quality

The results for this section are summarised in table 4. The results show that the preferences for different component choices vary between speakers.

The differences between the voices can be perceived and in some cases, preferences which show increase in output quality are shown across all participants listening to the voices.

Participant	Energy	Durations	GV for mel cep	Data
Speaker 1	Ave	Speaker	No diff.	No diff.
Listener 1A	Ave	Ave	No diff.	No diff.
Listener 1B	Speaker	Ave	No diff.	Unedited
Speaker 2	Ave	No diff.	Speaker	Unedited
Listener 2A	Ave	Ave	Ave	No diff.
Listener 2B	Ave	Ave	Ave	Edited
Listener 3A	No diff.	Ave	Ave	Edited

Table 4: *Preferences for quality shown by speakers and listeners for output synthesised in two difference conditions: speaker or average voice energy, duration or spectral global variance and using unedited or edited data. All other features remained constant.*

5.4. Appropriateness for speaker

The results for this question are shown in table 5 for output rate and global variance for log F0. The results show that differences between the conditions are discernible and preferences can be made for both rate of utterance and global variance for log F0 for speakers 1 and 2 but not for speaker 3.

5.5. Attitudes

For the rating of acceptability of the voice, from 1 (do not like the voice) - 5 (like the voice), speaker 1 rated his output voice as 5. He stated that he would be happy to use that voice on days when his own was not clear and showed no preference between the choice of using his own reconstructed voice or the Acapela voice. Speaker 2 rated his output voice as 1. He stated that he would not want to use that voice on days when his own was not clear and showed a preference for the Acapela voice over the presented reconstructed versions of his own voice.

	rates of speech			GV for log F0		
Participant	ave	slow	none	ave	sp.	none
Speaker 1	0	0	3	0	2	1
Listener 1A	2	1	0	3	0	0
Listener 1B	0	2	1	2	1	0
Total	2	3	4	5	3	1
Speaker 2	2	0	1	2	0	1
Listener 2A	1	1	1	2	1	0
Listener 2B	2	0	1	3	0	0
Total	5	1	3	7	1	1
Listener 3A	0	0	3	1	0	2

Table 5: *Number of utterances out of 3 preferred for different prosodic alterations. The conditions compared average voice durations (ave) and average voice durations slowed down (slow). The second experiment compared average voice gv-lF0 (ave) and the speaker's own gv-lF0 (sp.). None indicates the participants had no preference.*

6. Discussion

This study reports on the ratings of synthetic voices created for three dysarthric speakers. For these speakers, no examples exist of their premorbid speech so evaluation of the techniques used to create the synthetic voice are qualitative in nature. The speakers and familiar listeners were asked to rate different synthetic voices.

6.1. Acceptability of synthetic voices

The evaluation suggests that 150 or 200 sentences is not sufficient to fully capture the likeness of the speakers' voices using an average voice model matched for sex, but not other factors such as accent or age. The synthetic voice for Speaker 1 was rated more positively in terms of similarity by listeners than the voices for Speakers 2 and 3. It may be that this higher rating is due to the larger number of sentences used to create the voice for Speaker 1. It is also the case that the quality of the recorded data used for speaker 3's voice was lower, and his speech was also more severely impaired than the other two speakers. The manual data selection process used to extract the most intelligible portion may have introduced more variable quality sections into the adaptation data.

The perceived presence of "American-sounding" features, which is likely to be from the US English average voice, was common for all the voice. In-

deed Speaker 2 suggested this prevented him from recognising himself in the output voices. Listeners 2A and 2B also made this observation and emphasised that the English quality conveyed in speaker 2's voice was important to display his character. Listener 3A stated that the voice sounded like an American version of speaker 3 and then emphasised that this was not his identity. It may be surmised that an average voice more similar to the target voice in terms of accent may lead to a synthetic voice more acceptable to the speaker.

The listeners for speakers 1 and 2 noted that there were sections of the output with a strong likeness to the voice of the speaker participants, usually at the syllable level and the US English influence on the voice made it sound disjointed and less like the speakers. For speaker 1, the listeners agreed that those sections that did sound like speaker 1 had captured his voice well. Listener participant 3A noted that the voice with speaker log F0 and mel cepstra sounded somewhere between a generic speech synthesiser voice and the speaker's voice.

6.2. Feature selection

The factors influencing the quality of the voice output were dependent on the speaker and the effects of dysarthria on their speech. Where there were apparently large perceptual differences between voices those which were thought to be of higher quality or higher intelligibility were rated highest. However, this was not always the case when listeners perceived a voice to more accurately represent a speaker. For instance, although most listeners preferred to voices for Speakers 1 and 2 which used average voice energies, one listener (3A) noted that more accent of the target speaker could be heard in the examples using their own energies. This suggested that the perception of identity in speech is complex but also that the cues relating to these will need to be traded to achieve an acceptable synthetic voice.

Listener 3A heard no difference between the speech with the energy taken from the average voice and the speech using the speaker's own energy, this difference could have been masked by the overall lower quality of the output for speaker 3.

Speaker 1 preferred the voice where his own durations were used as he identified his own voice clearly in that example. Listeners 1A and 1B noted that that example sounded more like speaker 1 but in their judgements noted that they preferred the example with the average voice durations because it

was more intelligible and therefore more suitable for use with a communication aid. This was also true for listener 3A who noted that the speaker's own durations contributed to the identity of the speaker but who preferred the average voice durations because of the perceptual reduction of the impairment in the voice if it was to be used in a communication aid.

For Speaker 2, the duration information was not very different to his original speech and although both listeners preferred the average voice durations, they did note that the two outputs were very similar. Speaker 2 noted that although for one particular voice, the global variance for spectral information from the average voice made the output clearer, he preferred the voice with his own global variance for spectral information. This output produced a slightly muffled percept but this preference could be related to the perceived softness in the voice quality that it introduced, which Speaker 2 noted was missing in other examples. This was also noted by the listeners but they chose the average voice example as they recognised the need for the output to be clear and intelligible. The preferences for the global variance for spectral features across all participants suggests that it positively contributes to the output synthesis quality for these speakers.

These results suggest that to retain speaker characteristics in the synthetic voice the duration distributions of the target speaker should be retained. However, this is a feature which is likely to be affected by the individual's speech impairment. Although is appears that using speaker's own durations did contribute to the identification of the speaker, it is clear that substitution should be made when the speaker durations vary greatly from those of the average voice to maintain intelligibility of the synthetic voice. To reduce the effect of this substitution it is likely that using an average voice with the same regional accent would be better suited to capture the duration aspects of the accent of the individual. If a choice of regionally-appropriate average voices were available, it would offer a more appropriate set of duration characteristics to more closely replicate the accent of the speaker.

The differences in output rate could be perceived by some of the participants, although there is a limited extent to which the rate can be slowed until it starts to reduce intelligibility, as observed during the informal listening tests conducted by the author. The rate of output was therefore only slowed slightly, which may not have been sufficient for all participants to observe. Where the difference was perceived, this contributed to the individual preferences along with observing where certain speaker characteristics were more strongly perceived in certain stimuli.

The change of global variance for log F0 could also be perceived by some of the participants. Speaker 2, who had a relatively narrow range of log F0 preferred to have a wider range than his own in the output. Speaker 1's range was closer to the average voice and the preference showed it was more appropriate for him. Where speaker 1 did notice a difference in what he heard for the global variance for log F0 stimuli, he said that the difference was that one was easier to understand than the other. This was supported by the evaluators who also used intelligibility to make their judgements but were also listening more closely to identify bits of speaker 1's accent and the Americanised output. Listener 3A identified a difference in the output of one stimulus, preferring the average voice global variance for log F0. The difficulty in recognising a difference between the stimuli for these parameter changes could be related to the overall quality of the output synthesis for this speaker, although it is difficult to draw conclusions based on the limited amount of results for this speaker.

In relation to the pathologies of the speakers, both 1 and 2 had variable energy in their speech and both preferred voices with normalised energy output. Speaker 2's monopitch output was reconstructed to have a preferred wider variability in pitch. This factor could also be altered for speaker 3 to widen the log F0 variability found in his speech data. Imprecise articulations present in all speakers' data were handled by using the average voice model durations and global variance for spectral features. Selecting data for adaptation also contributed to the reconstruction quality, particularly for the more severely impaired speech of speaker 3.

In terms of acceptability, this evaluation shows that different people have different priorities for their VOCA use and this highlights the need to provide more choice and more customisation for voices that are provided with communication aids to fit the wants and needs of individuals. The evaluation also provided insight into the importance of some individuals' voices to them as their marker of identity. The reactions within the evaluation suggested that if a voice is said to be personalised to match that of an individual, then the point of acceptability of that voice reconstruction is dependent on the individual. What is also clear is that if the user does not accept the voice then they do not want it to represent them, again supporting the case that customisation, choice and adaptation to the individual is important for the acceptability of such devices.

7. Conclusion

Using HMM-based synthesis with data selection and imposition of information from the average voice model is a promising technique to reconstruct voices of these individuals with mild to moderate dysarthria. These results point to more success being achieved and better similarity judged if the American influence on the voices was removed. Using an average voice that is more appropriate to the speaker's own accent would reduce the difference between the speaker characteristics of the average voice and the adaptation data. This would reduce the perceptions that were reported of that was apparent in the synthesised output in the evaluation. This led to the percept of hearing more than one speaker in the voice as noted by all the listener participants. Speakers with dysarthria find it more difficult to produce the amount of data needed to fully adapt all the characteristics contained in the average voice to their own. It is hypothesised that the use of a more regionally-appropriate British English average voice model would improve this process for this amount of data. Since the experiments reported here, ongoing work with building HTS voices with British English data means that UK average voice models are now available along with multi-accented English speaking average voices (Yamagishi et al., 2008b). Further work also points to the use of starting point voices which are closer to the target produce higher quality output (Yamagishi et al., 2010).

8. Acknowledgements

STRAIGHT is used within HTS with permission from Hideki Kawahara. Sarah Creer's PhD work was funded by the Engineering and Physical Sciences Research Council (EPSRC), UK.

References

Anastasakos T, McDonough J, Schwartz R, Makhoul J. A compact model for speaker adaptive training. In: Proceedings of ICSLP. 1996. p. 1137–40. Philadelphia: PA, USA.

Angelo DH, Kokosa SM, Jones SD. Family perspective on augmentative and alternative communication: families of adolescents and young adults. Augmentative and Alternative Communication 1996;12(1):13–20.

von Bekesy G. The structure of the middle ear and the hearing of one's own voice by bone conduction. Journal of the Acoustical Society of America 1949;21:217–32.

Bunnell HT, Lilley J, Pennington C, Moyers B, Polikoff J. The ModelTalker system. In: Proceedings of the Blizzard Challenge Workshop. 2010. Kansai Science City, Japan.

Carmichael J, Green P. Devising a system of computerised metrics for the Frenchay Dysarthria Assessment intelligibility tests. In: Proceedings of the University of Cambridge First Postgraduate Conference in Language Research: CAMLING. Cambridge; 2003. p. 473–9.

Chambers JK. Sociolinguistic Theory. Oxford: Blackwell, 1995.

Crabtree M, Mirenda P, Beukelman DR. Age and gender preferences for synthetic and natural speech. Augmentative and Alternative Communication 1990;6(4):256–61.

Creer SM, Green PD, Cunningham SP, Yamagishi J. Building personalised synthetic voices for individuals with dysarthria using the HTS toolkit. In: Mullennix JW, Stern SE, editors. Computer Synthesised Speech Technologies: Tools for Aiding Impairment. Hershey, PA, USA: IGI Global; 2010. p. 92–115.

Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. Journal of Speech and Hearing Research 1969;12:246–69.

van Dommelen WA. Acoustic parameters in human speaker recognition. Language and Speech 1990;33(3):259–72.

Duffy J. Motor speech disorders: substrates, differential diagnosis and management. 2nd ed. St Louis, MO: Elsevier Mosby, 2005.

Enderby P, Emerson L. Does speech and language therapy work? London: Whurr, 1995.

Enderby PM. Frenchay Dysarthria Assessment. Austin, TX: Pro-ed, 1983.

Fujisaki H, Hirose K. Analysis of voice fundamental frequency contours for declarative sentences of japanese. Journal of the Acoustical Society of Japan (E) 2000;5(4):233–42.

Furui S. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech and Signal Processing 1981;29(2):254–72.

Hetzroni OE, Harris OL. Cultural aspects in the development of AAC users. Augmentative and Alternative Communication 1996;12(1):52–8.

Holmberg E, Nordqvist K, Ahlström G. Prevalence of dysarthria in adult myotonic dystrophy (m. steinert) patients; speech characteristics and intelligibility. Logopedics Phoniatrics Vocology 1996;21(1):21–7.

Huckvale M. SCRIBE manual v1.0. http://phon.ucl.ac.uk/resource/scribe/scribe-manual.htm; 2004. Last accessed 18 August 2009.

Huckvale M. Prorec 1.2. Software; 2009. Available at: http://www.phon.ucl.ac.uk/resource/prorec, Last accessed 24 March 2009.

Isogai J, Yamagishi J, Kobayashi T. Model adaptation and adaptive training using ESAT algorithm for HMM-based speech synthesis. In: Proceedings of Eurospeech. 2005. p. 2597–600. Lisbon, Portugal.

Jreige C, Patel R, Bunnell HT. Vocalid: Personalizing text-to-speech synthesis for individuals with severe speech impairment. In: Proceedings of ASSETS. 2009. p. 259–60. Pittsburgh: PA, USA.

Katz WF. Anticipatory coarticulation and aphasia: implications for phonetic theories. Journal of Phonetics 2000;28(3):313–34.

Kawahara H, Masuda-Katsuse I, de Cheveigné A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Communication 1999;27:187–207.

Kominek J, Black AW. CMU Arctic databases for speech synthesis. http://festvox.org/cmu_arctic/cmu_arctic_report.pdf; 2003. Last accessed 20 April 2006.

Lasker JP, Bedrosian JL. Promoting acceptance of augmentative and alternative communication by adults with acquired communication disorders. Augmentative and Alternative Communication 2001;17(3):141–53.

Light J. Interaction involving individuals using augmentative and alternative communication systems: state of the art and future directions. Augmentative and Alternative Communication 1988;4(2):66–82.

Light J, Page R, Curran J, Pitkin L. Children's ideas for the design of AAC assistive technologies for young children with complex communication needs. Augmentative and Alternative Communication 2007;23(4):274–87.

Mathy P, Yorkston KM, Gutmann ML. AAC for individuals with amyotrophic lateral sclerosis. In: Beukelman DR, Yorkston K, Reichle J, editors. Augmentative communication for adults with neurogenic and neuromuscular disabilities. Baltimore: MD, Paul H. Brookes; 2000. p. 183–229.

Miller N, Noble E, Jones D, Burn D. Life with communication changes in Parkinson's disease. Age and Ageing 2006;35:235–9.

Murphy J. 'I prefer contact this close': perceptions of AAC by people with motor neurone disease and their communication partners. Augmentative and Alternative Communication 2004;20(4):259–71.

Murray IR, Arnott JL. A tool for the rapid development of new synthetic voice personalities. In: Speech and Language Technology for Disabled Persons. 1993. p. 111–4. Stockholm, Sweden.

O'Keefe BM, Brown L, Schuller R. Identification and rankings of communication aid features by five groups. Augmentative and Alternative Communication 1998;14(1):37–50.

Parette P, Huer MB. Working with Asian American families whose children have augmentative and alternative communication needs. Journal of Special Education Technology E-Journal 2002;17(4). Http://jset.unlv.edu/17.4T/parette/first.html, last accessed 25 October 2006.

Shinoda K, Watanabe T. MDL-based context-dependent subword modelling for speech recognition. Journal of the Acoustical Society of Japan (E) 2000;21:79–86.

Smith MM. The dual challenges of aided communication and adolescence. Augmentative and Alternative Communication 2005;21(1):76–9.

Toda T, Tokuda K. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Transactions on Information and Systems 2007;E90-D(5):816–24.

Tokuda K, Yoshimura T, Masuko T, Kobayashi T, Kitamura T. Speech parameter generation algorithm for HMM-based speech synthesis. In: Proceedings of ICASSP. 2000. p. 1315–8. Beijing, China.

Weismer G. Motor Speech Disorders. Oxford: Plural Publishing, 2007.

Wells JC. Accents of English: An Introduction. Cambridge: Cambridge University Press, 1982.

Yamagishi J, Kobayashi T, Nakano Y, Ogata K, Isogai J. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE transactions on Audio, Speech and Language Processing 2009;17(1):66–83.

Yamagishi J, Ling Z, King S. Robustness of HMM-based speech synthesis. In: Proceedings of Interspeech. 2008a. p. 581–4. Brisbane, Australia.

Yamagishi J, Veaux C, King S, Renals S. Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction. Acoustical Science and Technology 2012;33(1):1–5.

Yamagishi J, Watts O, King S, Usabaev B. Roles of the average voice in speaker-adaptive HMM-based speech synthesis. In: Proceedings of Interspeech. 2010. p. 418–21. Makuhari, Japan.

Yamagishi J, Zen H, Toda T, Tokuda K. Speaker-independent HMM-based speech synthesis system – HTS-2007 for the Blizzard challenge 2007. In: Proceedings of the Blizzard Challenge Workshop. 2007. Bonn, Germany.

Yamagishi J, Zen H, Wu YJ, Toda T, Tokuda K. The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard challenge. In: Proceedings of the Blizzard Challenge Workshop. 2008b. Brisbane, Australia.

Young S, Everman G, Kershaw D, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P. The HTK book version 3.2.1; 2002. .

Zen H, Nose T, Yamagishi J, Sako S, Masuko T, Black AW, Tokuda K. The HMM-based speech synthesis system (HTS) version 2.0. In: Proceedings of the 6th International Workshop on Speech Synthesis. 2007. p. 294–9. Bonn, Germany.

Zen H, Toda T. An overview of Nitech HMM-based speech synthesis system for Blizzard challenge 2005. In: Proceedings of Interspeech. 2005. p. 93–6. Lisbon, Portugal.

Zen H, Tokuda K, Black A. Statistical parametric speech synthesis. Speech Communication 2009;51(11):1039–64.

Zen H, Tokuda K, Masuko T, Kobayashi T, Kitamura T. Hidden semi-Markov model based speech synthesis. In: Proceedings of ICSLP. 2004. p. 1397–400. Jeju Island, South Korea.

Ziegler W. Apraxia of speech. In: Handbook of Clinical Neurology. Amsterdam: Elsevier Science; volume 88 of *Neuropsychology and behavioural neurology*; 2008. p. 269–86.